BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e35626
doi: 10.3897/biss.3.35626

OPEN ACCESS

Conference Abstract

Data Quality Task Group 2: Tests and Assertions

Lee Belbin[‡], Arthur D Chapman[§], John Wieczorek[|], Paula Zermoglio[¶], Paul J Morris[#]

‡ Atlas of Living Australia, CSIRO, Canberra, Australia
§ Australian Biodiversity Information Services, Ballan, Australia
| Museum of Vertebrate Zoology, University of California, Berkeley, United States of America
¶ Instituto de Ecología, Genética y Evolución de Buenos Aires (IEGEBA-CONICET), University of Buenos Aires, Buenos Aires, Argentina
# Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America

Corresponding author: Lee Belbin (leebelbin@gmail.com)

Received: 22 Apr 2019 | Published: 10 Jul 2019

Citation: Belbin L, Chapman AD, Wieczorek J, Zermoglio P, Morris PJ (2019) Data Quality Task Group 2: Tests and Assertions. Biodiversity Information Science and Standards 3: e35626. https://doi.org/10.3897/biss.3.35626

Abstract

'Data Quality Test and Assertions' Task Group 2 (https://www.tdwg.org/community/bdq/tg-2/) has taken another year to clarify the 102 tests (https://github.com/tdwg/bdq/issues?q=is%3Aissue+is%3Aopen+label%3ATest). The original mandate to develop a core suite of tests that could be widely applied from data collection to user evaluation of aggregated data seemed straight-forward. Two years down the track, we have proven that to be incorrect. Among the final tests are complexities that none of the core group anticipated. For example, the need for a definition of 'empty' or the 'Expected response' from the test under various scenarios.

The record-based tests apply to Darwin Core terms (https://dwc.tdwg.org/terms/) and have been classified as of type validation (66), amendment (29), notification (3) or measure (5). Validations test one or more Darwin Core terms against known characteristics, for example, VALIDATION_MONTH_NOTSTANDARD. Amendments may be applied to Darwin Core terms where we can unambiguously offer an improvement to the record, for example, AMENDMENT_MONTH_STANDARDIZED. Notifications are made where we believe a flag will help alert users to an issue that needs evaluation, for example, NOTIFICATION_DATAGENERALIZATIONS_NOTEMPTY. Measures are summaries of test outcomes at the record level, for example, MEASURE_AMENDMENTS_PROPOSED.

We note that 41 require some parameters to be established at the time of test implementation, 20 tests require access to a currently accepted vocabulary and 3 tests rely on ISO/DCMI standards. The dependency on vocabularies to circumscribe permissible values for Darwin Core terms led to the establishment by Paula Zermoglio of DQ Task Group 4 (https://github.com/tdwg/bdq/tree/master/Vocabularies). A vocabulary of 154 terms that are associated with the tests and assertions have been developed.

As at the time of writing this abstract, test data and demonstration code implementation of each test are yet to be completed. We hope these will be finalized by the time of this presentation.

Keywords

Fitness for use, fitness for purpose, tests, Darwin Core, quality.

Presenting author

Lee Belbin